Bayesian statistical analysis of hydrogeochemical data using point processes: a new tool for source detection in multicomponent fluid mixtures

Christophe Reype, Antonin Richard, Madalina Deaconu, Radu S. Stoica

Bayesian statistical analysis of hydrogeochemical data using point processes: a new tool for source detection in multicomponent fluid mixtures

C. Reype[1], A. Richard[2], M. Deaconu[1], and R. S. Stoica[3]

[1] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France*
[2] *Université de Lorraine, CNRS, GeoRessources, F-54000 Nancy, France*
[3] *Université de Lorraine, CNRS, IECL, F-54000 Nancy, France*

September 2020

Abstract

Hydrogeochemical data may be seen as a point cloud in a multi-dimensional space. Each dimension of this space represents a hydrogeochemical parameter (*i.e.* salinity, solute concentration, concentration ratio, isotopic composition...). While the composition of many geological fluids is controlled by mixing between multiple sources, a key question related to hydrogeochemical data set is the detection of the sources. By looking at the hydrogeochemical data as spatial data, this paper presents a new solution to the source detection problem that is based on point processes. Results are shown on simulated and real data from geothermal fluids.

Introduction

The composition of many geological fluids is controlled by variable contributions of multiple sources (*e.g.* seawater, meteoric water, hydrothermal water). The knowledge of these sources helps to built conceptual and quantitative models of fluid and mass transfer in the Earth's crust (Yardley & Bodnar, 2014). If the sources are known, the contribution of each sources in every mixture can be inferred from hydrogeochemical data (*e.g.* Carrera, Vázquez-Suñé, Castillo, & Sánchez-Vila, 2004; Skuce, Longstaffe, Carter, & Potter, 2015). In the case where the sources are not known, they can be inferred from the data (*e.g.* Pinti et al., 2020).

The paper presents a Bayesian method of source detection based on point processes. The method is inspired by pattern detection methodologies used in image analysis, animal epidemiology and astronomy (R. Stoica, Descombes, & Zerubia, 2004; R. S. Stoica, Gay, & Kretzschmar, 2007; R. S. Stoica, Martinez, Mateu, & Saar, 2005; R. S. Stoica, Martínez, & Saar, 2007).

1 Materials and methods

Let $\mathbf{x} = \{x_i, i = 1, \ldots, n\}$ be a set of sources, giving the source position within a multi-dimensional (in practice two-dimensional) space formed by the hydrogeochemical parameters. A data point d is a mixture of these sources (*i.e.* it is explained by these sources) if it is a barycenter of these sources as stated in (Faure, 1997) *i.e.*

$$d = \sum_{i=1}^{n} \gamma_i x_i \tag{1}$$

with $0 \leq \gamma_i \leq 1$ for each i and $\sum_{i=1}^{n} \gamma_i = 1$.

In the Euclidean plane, the source pattern (*i.e.* set of sources) is unknown and also somehow outlined by the set of hydrogeological data points. The key hypothesis at the basis of our work is that this pattern is made of interacting points. A preliminary condition for our model is that the hidden sources pattern exhibits the following properties :

- the number of sources is not known but it should be controlled or minimal in a certain sense

- two sources cannot be too close
- the data points originating from a mixture of sources should be rather close to them
- the convex hull enclosing the set of data points is enclosed within the convex hull given by the source positions.

These hypotheses allow to consider the sources as a realisation of a point process described by a Gibbs probability density:

$$p(\mathbf{x}|\theta) = \frac{\exp\left[-U(\mathbf{x}|\theta)\right]}{Z(\theta)} = \frac{\exp\left[-U_{\mathbf{d}}(\mathbf{x}|\theta) - U_i(\mathbf{x}|\theta)\right]}{Z(\theta)}$$

with $\mathbf{x}$ the configuration of sources (or set of sources), $Z(\theta)$ the normalising constant and U the energy function.

The energy function is built as the sum of two components. The first term, $U_{\mathbf{d}}(\mathbf{x}|\theta)$ is the data term and it controls the positioning of the sources with respect to the observed data points $\mathbf{d} = \{d_1, d_2, \ldots, d_m\}$. Its expression is given by

$$U_{\mathbf{d}}(\mathbf{x}, \theta) = \theta_1 g(\mathbf{x}, \mathbf{d}) + \theta_2 \sum_{j=1}^{m} \alpha(d_j, \mathbf{x}) + \theta_3 n_e(\mathbf{x}, \mathbf{d}).$$

Here $g(\mathbf{x}, \mathbf{d})$ is the absolute value difference between the area of the sources and the data point convex hull, respectively. The function $\alpha(d_j, \mathbf{x})$ represents the minimum distance between the data point d_j and the sources cloud and it is given by $\min\{\|d_j - x_i\|_2^2 : i = 1, \ldots, n(\mathbf{x})\}$, with $n(\mathbf{x})$ the number of sources in the configurations. The measure $n_e(\mathbf{x}, \mathbf{d})$ counts the number of data points in $\mathbf{d}$, that belong to the convex hull given by $\mathbf{x}$. The parameters $\theta_1, \theta_2 \geq 0$ and $\theta_3 \leq 0$ are chosen such that to penalize important differences between the convex hull areas, to encourage the source to be situated rather close to the data points and to increase the number of data points that are explained by the sources, respectively.

The second term, $U_i(\mathbf{x}|\theta)$ is the interaction term and it writes as

$$U_i(\mathbf{x}, \theta) = \theta_4 n(\mathbf{x}) + \theta_5 n_r(\mathbf{x}).$$

with $n_r(\mathbf{x})$ the number of pairs of sources at distance shorter than r, which is a pre-fixed known value. The parameters $\theta_4, \theta_5 \geq 0$ are chosen in order to penalize a too high number of sources and pairs of sources situated too close, respectively.

The point process on a finite domain W (*i.e.* $\mu(W) = \int_W \xi \, d\xi < \infty$), that is defined by the previous energy function, is well defined and locally stable R. Stoica (2014). Based on these terms, the model is able to generate point configurations that exhibit the properties required by the assumed hypotheses. The source pattern is estimated by the point configuration that maximises the probability density $p(\mathbf{x}|\theta)$

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\Omega} p(\mathbf{x}|\theta) = \arg\min_{\mathbf{x}\in\Omega} U(\mathbf{x}|\theta). \tag{2}$$

The solution of the problem (2) is obtained by implementing a simulated annealing algorithm. This algorithm is a global optimisation method that iteratively samples from $p(\mathbf{x}|\theta)^{1/T}$ while making $T \to 0$ slowly. Convergence properties of this algorithm are shown in R. S. Stoica, Gregori, and Mateu (2005).

1.1 Optimisation algorithm

The implemented simulated annealing algorithm has the following structure:

1) set θ, T, k_{max}, $\mathbf{x}^{(0)}$, c and $k = 1$
2) while $k \leq k_{max}$

a) generate $\mathbf{x}^{(k)}$ with probability $p(\mathbf{x}^{(k-1)}|\theta)^{1/T}$

b) set $T = c * T$ and $k = k + 1$

3) set $\mathbf{x} = \mathbf{x}^{(k)}$

This structure implements a sub-optimal cooling schedule, for practical reasons. An optimal logarithmic cooling schedule as specified by R. S. Stoica, Gregori, and Mateu (2005) may be considered.

The sampling of $p(\mathbf{x}|\theta)$ is done via the Metropolis-Hasting algorithm described below :

1) set r_c, p_b, p_d, p_c with $p_b + p_d + p_c \leq 1$

2) with probability p_b choose birth, with probability p_d choose death and with probability p_c choose change.

birth: a) generate a random point η on W and set $\mathbf{x}' = \mathbf{x} \cup \{\eta\}$

b) calculate $\beta_b = \min\{1, \frac{p_d}{p_b} \frac{p(\mathbf{x}\cup\{\eta\}|\theta)}{p(\mathbf{x}|\theta)} \frac{\mu(W)}{n(\mathbf{x})+1}\}$

death: a) choose a point η of $\mathbf{x}$ and set $\mathbf{x}' = \mathbf{x} \setminus \{\eta\}$

b) calculate $\beta_d = \min\{1, \frac{p_b}{p_d} \frac{p(\mathbf{x}\setminus\{\eta\}|\theta)}{p(\mathbf{x}|\theta)} \frac{n(\mathbf{x})}{\mu(W)}\}$

change: a) choose a point η of $\mathbf{x}$ and generate a random point ξ in the ball $B(\eta, r_c)$ and set $\mathbf{x}' = \mathbf{x} \setminus \{\eta\} \cup \{\xi\}$

b) calculate $\beta_c = \min\{1, \frac{p(\mathbf{x}\setminus\{\eta\}\cup\{\xi\}|\theta)}{p(\mathbf{x}|\theta)}\}$

3) the new configuration $\mathbf{x} = \mathbf{x}'$ is accepted with the appropriate probability β ; otherwise the algorithm remains in the same state $\mathbf{x}$.

The previous dynamic is $\phi-$ irreducible, Harris recurrent and geometric ergodic, guaranteeing the convergence of the algorithm towards the distribution of interest given by $p(\mathbf{x}|\theta)$ Moller and Waagepetersen (2003); R. Stoica (2014); van Lieshout (2000).

2 Results

The proposed model was tested on two different data sets. The first one is a simulated data set, the second one is a set of hydrogeochemical data from geothermal fluids described in (Pinti et al., 2020). The model is coded in C++, and the results are displayed in R with the library "ggplot".

We set $T = 1000$, $k_{max} = 10000$, $c = 0.995$, $r_c = 0.3$, $p_b = 0.35$, $p_d = 0.35$ and $p_c = 0.3$. The parameters θ were chosen separately, for each data set, after several trials and errors.

2.1 Simulated data

The data are created by generating three sources. The vector of contributions of each sources to a data point is generated by a Dirichlet law with parameters $(1, 1, 1)$. Hence the data points are points uniformly distributed in the convex hull given by the sources positions in the 2D space of hydrogeochemical parameters. Moreover, a Gaussian noise, of mean 0 and variance 10^{-1} for each coordinate, was added to each datapoint to represent the noise during the measurement.

We set the parameters to $\theta = (100, 1, -100, 700, 50)$ and $r = 2$. The results are shown in Figure 1. The black dots are the data points, the blue symbols are the real sources and the gradient of blue color shows the density of simulated sources.

There are three areas that exhibit a high density of simulated sources, which corresponds to the actual number of sources. Moreover their positions are really close to the real sources.

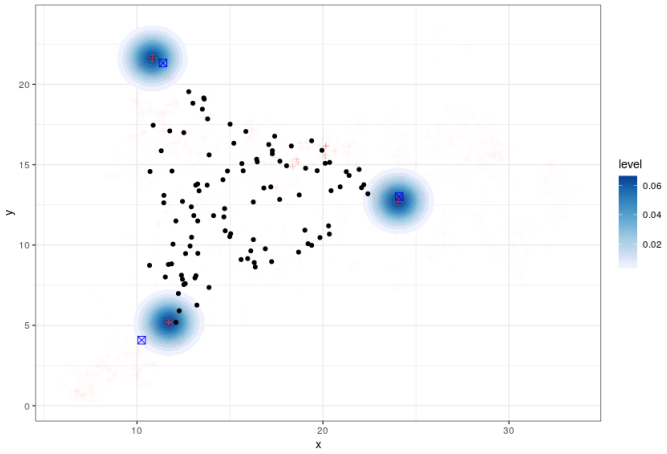

Figure 1: *Point process source detection for the simulated data in the case of a three-component fluid mixing system where x and y are the concentrations of two solutes (arbitrary units)*

2.2 Real data

We are now comparing the results of our model with the results of the model presented in (Pinti et al., 2020). This model gives the smallest triangle (in term of area) that enclose the data. In Figure 2 we set $\theta = (200, 1, -1200, 210, 5)$, and in Figure 3 we set $\theta = (196, 2, -1200, 220, 5)$. The parameters are completely different than in 1 because the data are not in the same scale.

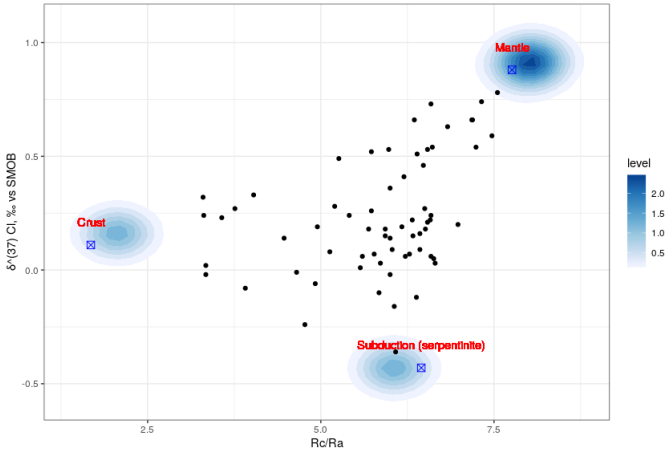

Figure 2: *Point process source detection for a three-component geothermal fluid mixing system where $\delta^{37}Cl$, $‰$vs SMOC and Rc/Ra are respectively the stable isotopic composition of chlorine and the $^4He/^3He$ ratio of the samples (data from (Pinti et al., 2020, Figure 4))*

The model is able to detect the number and the position of the sources inferred in (Pinti et al., 2020), while the model sufficient statistics provides a more complete morpho-statistical description of the sources.

3 Conclusions and perspectives

Clearly, the use of this method requires at least partial knowledge regarding the model parameters. Such a knowledge is built by embedding the available geological information into prior distributions.

This new tool should be improved in order to become helpful not only in the analysis of geological

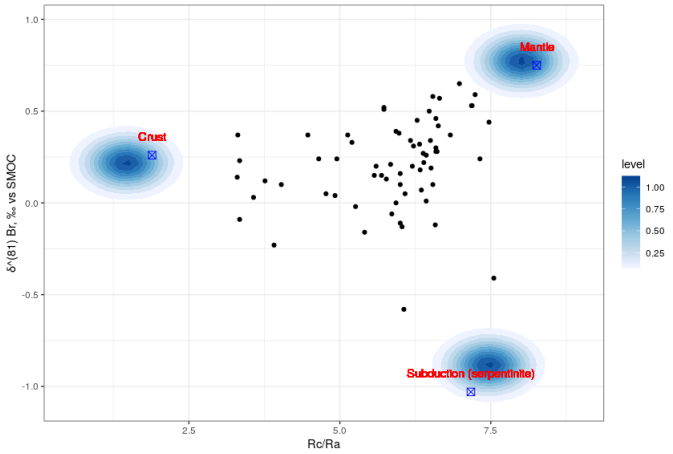

Figure 3: *Point process source detection for a three-component geothermal fluid mixing system where $\delta^{81}Br$, $‰$vs SMOB and Rc/Ra are respectively the stable isotopic composition of bromine and the $^4He/^3He$ ratio of the samples (data from (Pinti et al., 2020, Figure 5))*

fluids but also in other fields that deals with mixtures (*e.g.* Longman et al., 2018; Phillips & Gregg, 2003). This is possible due to the use of the embedded spatial and Bayesian paradigms.

Acknowledgments

This work was performed in the frame of the DEEPSURF project (http://lue.univ-lorraine.fr/fr/impact-deepsurf) at Université de Lorraine. This work was supported partly by the french PIA project Lorraine Université d'Excellence, reference ANR-15-IDEX-04-LUE.

References

Carrera, J., Vázquez-Suñé, E., Castillo, O., & Sánchez-Vila, X. (2004). A methodology to compute mixing ratios with uncertain end-members. *Water resources research*, *40*(12).

Faure, G. (1997). *Principles and applications of geochemistry* (Vol. 625). Prentice Hall New Jersey, United States,.

Longman, J., Veres, D., Ersek, V., Phillips, D. L., Chauvel, C., & Tamas, C. G. (2018). Quantitative assessment of pb sources in isotopic mixtures using a bayesian mixing model. *Scientific reports*, *8*(1), 6154.

Moller, J., & Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC.

Phillips, D. L., & Gregg, J. W. (2003). Source partitioning using stable isotopes: coping with too many sources. *Oecologia*, *136*(2), 261–269.

Pinti, D. L., Shouakar-Stash, O., Castro, M. C., Lopez-Hernández, A., Hall, C. M., Rocher, O., ... Ramírez-Montes, M. (2020). The bromine and chlorine isotopic composition of the mantle as revealed by deep geothermal fluids. *Geochimica et Cosmochimica Acta*.

Skuce, M., Longstaffe, F., Carter, T., & Potter, J. (2015). Isotopic fingerprinting of groundwaters in southwestern ontario: Applications to abandoned well remediation. *Applied Geochemistry*, *58*, 1–13.

Stoica, R. (2014). Modélisation probabiliste et inférence statistique pour lanalyse des données spatialisées. *Research Habilitation Thesis, Université Lille*, *1*.

Stoica, R., Descombes, X., & Zerubia, J. (2004). A gibbs point process for road extraction from remotely sensed images. *International Journal of Computer Vision*, *57*(2), 121–136.

Stoica, R. S., Gay, E., & Kretzschmar, A. (2007). Cluster pattern detection in spatial data based on monte carlo inference. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *49*(4), 505–519.

Stoica, R. S., Gregori, P., & Mateu, J. (2005). Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, *115*, 1860-1882.

Stoica, R. S., Martinez, V. J., Mateu, J., & Saar, E. (2005). Detection of cosmic filaments using the candy model. *Astronomy & Astrophysics*, *434*(2), 423–432.

Stoica, R. S., Martínez, V. J., & Saar, E. (2007). A three-dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *56*(4), 459–477.

van Lieshout, M. N. M. (2000). *Markov point processes and their applications*. Imperial College Press, London.

Yardley, B. W., & Bodnar, R. J. (2014). Fluids in the continental crust. *Geochemical Perspectives*, *3*(1), 1–2.